EnviroInfo 2004 (Geneva)
Sh@ring – EnviroInfo 2004
 ISBN: 28 29 30 275-3

US EPA's System of Registries

Larry Fitzwater[1]

1. Metadata and EPA's System of Registries (SoR)

1.1 Why metadata?

Simply stated, metadata is data about data. Metadata provides context to data allowing users to understand, locate, retrieve, and use it appropriately. In the presence of well-described metadata, data becomes information.

Metadata:

- Gives data meaning, relevance and purpose
- Specifies its structure, content, and context
- Aids in discovery of relevant information
- Helps facilitate system interoperability and data integration

Examples of metadata are:

- Ingredients and nutritional information on food packaging
- The date and time an email was delivered
- A library's reference information for books
- Time, location, and methods used in collecting water samples
- The name, date, and time associated with an electronic file

Why is metadata important? Metadata allows users to identify data - imagine what would happen if your PC randomly assigned names to your files rather than allowing you to assign file names. Metadata also helps users understand data - a string of 10 digits such as 2854697102 is ambiguous without additional information indicating it's a telephone number.

1.2 Why EPA developed a System of Registries (SoR)

EPA uses numerous registries (physical database tables) and Web interfaces to manage and make metadata available. The SoR is a Web interface that supports accessing and using the following core registries:

- Environmental Data Registry (EDR)
- Facility Registry System (FRS)

[1] US Environmental Protection Agency, 1200 Pennsylvania Avenue NW, MC 2822T, Washington, DC 20460 Phone: 202-566-1653 E-Mail: fitzwater.larry@epa.gov

- Registry of EPA Applications and Databases (READ)
- Substance Registry System (SRS)
- XML Registry
- Environmental Management Information System (EIMS)
- Terminology Reference System

The SoR combines both fully integrated systems (EDR, READ, SRS, XML Registry, and TRS) and linked applications (EIMS, an Office of Research and Development product, and FRS, an Office of Environmental Information (OEI) product).

The SoR contains:

- Data elements
- eXtensible Markup Language (XML) tags and schema
- Data standards
- Regulated substances
- Regulated facilities
- Environmental terms
- Information resource metadata
- Regulations, and
- Datasets that the Agency uses in its core business processes

The SoR was developed to support the Agency's data standards program and metadata management needs. These, in turn, support numerous Agency information technology initiatives, including the Agency enterprise architecture and data exchange with stakeholders through network

The SoR represents an ongoing effort to better manage Agency information resources as strategic assets in support of sound environmental decision making. This effort was initially conceptualized in the early 1990's in response to numerous Office of Management and Budget, General Accounting Office, and Inspector General reports calling for improving longstanding information technology weaknesses. At the time, the Agency was described as being "data rich but information poor."

In 1997, EPA responded by initiating an Agency-wide, strategic information management planning process that included an Information and Data Management Program focused on establishing data management policies and standards to improve and maintain data integrity. In 2003, the SoR became an integral component to EPA's enterprise architecture planning effort.

1.3 How registries support information management at EPA

The following are examples of how the registries support information management at EPA:

- The Environmental Data Registry helps improve data consistency by reducing duplication associated with redundant storage and maintenance, and by making well-formed data elements available for reuse.

- The Environmental Data Registry supports EPA's data standards program by storing standards metadata, providing access to standards, by and providing a tool that supports harmonizing of program metadata for standards implementation.
- The Substance Registry System reduces the need for redundant storage and ensures data quality by providing an authoritative resource for EPA regulated substances.
- The Facility Registry System serves as a backbone for Envirofacts (EPA's principal public access database) by ensuring that facilities are fully described and not double counted.
- The XML Registry serves XML schema to the Exchange Network by ensuring that states use approved schema that incorporate data standards, which will improve the quality of data coming into EPA.

1.4 Who the SoR serves and how it can be used

In addition to addressing the needs of EPA, the SoR is intended to address the needs of a diverse audience including:

- State and Tribal partners of the Exchange Network
- Industry
- Associations
- Non-profits
- Educational institutions
- Internal EPA customers

The following are examples of ways the SoR might be used:

- Information system managers and developers may be interested in data standards, XML, and application metadata. Tools within the EDR are available to support database harmonization and integration.
- Regulation writers can use data standards to shape data requirements for information collection on permit applications. They can read the Working for You page to see how data standards can be incorporated into regulations.
- Industry and the public may be interested in regulated substances located in the SRS, environmental terms defined in the TRS, or environmental information related to facilities found in the FRS.
- State and local agencies may want to follow the progress and status of Agency data standards in the EDR.
- System developers can use the EDR to download code sets for reuse.
- Federal government agencies might utilize the EDR metamodel as an off-the-shelf solution for their own metadata management initiatives.
- Environmental decision makers might review the *Registry Update* Newsletter to find out about EPA's data standards program and registry-related developments.

- Researchers can search for descriptive information about scientific environmental data sets in the EIMS.

1.5 How applications can dynamically link to the SoR

The SoR includes a stand-alone application known as the Environmental Metadata Gateway (EMG) portal that provides customized access to information in the registries. EMG includes multiple search engine portals that publish information from the underlying SoR metadata in EPA default or user-defined look-and-feel templates. This enables users, both inside and outside EPA, to use a URL to search and seamlessly navigate to pages displaying metadata registry content.

At this time, the EMG portal is being used to serve standard data element metadata to the Web site of the Environmental Data Standards Council and XML metadata to the Environmental Exchange Network XML Registry. In the future, EMG will expand with integrated search capabilities and the ability to transfer information to and from the registries in XML through an XML Gateway feature.

Go to Working for You to see how EPA's Office of Air and Radiation (OAR) partnered with OEI to dynamically link with the registries to support the Radiation Information System for Cleanup Sites (RISCS) repository.

1.6 How data quality and data integrity are maintained within the SoR

Data elements in the EDR, terms in the TRS, and substances in the SRS are registered in association with an information resource, which might be an application system, a standard document, a thesaurus, a law, or a regulation. The information resources are loaded into the database as classification schemes, which enable hierarchical organization of information within schemes. In this way, the registries document systems, their structure, and the semantic meaning of their contents.

All data loaded into the registries is dated upon entry and update. In conformance with the ISO/IEC 11179 model, the database supports versioning of all objects.

Data standards and applications can be registered with new versions. Permissible values associated with data elements are registered with a beginning and end date, so that changes can be tracked in code set lists.

Application metadata is considered to be high quality when it accurately represents what is stored in the application documentation. To help ensure data quality, most data loaded into the registries have undergone peer/quality assurance review. Data standard metadata is held to ISO/IEC 11179 standards for naming, definition, and completeness.

The registry design is primarily based on ISO/IEC 11179. ISO/IEC 11179-3:2003 is the international standard that defines the metamodel for metadata registries. Its

adoption is significant because it enables other government agencies to use EPA's registry as an off-the-shelf solution for developing future registries. The Veteran's Administration (VA) has successfully used EPA's registry to do this, and the benefits to both agencies have been significant. VA has leveraged EPA's experience and our metadata solution; EPA in turn has obtained additional functionality from VA registry buildouts.

Some extensions to the standard metamodel have been made - for example, information resources are documented according to the Dublin Core standard. Other model extensions have been made to support EPA/State standards for storage of facility, chemical, biological, location, and other data.

1.7 Standards and practices used in registry development and maintenance

The success and the utility of the SoR are based on several standards and best practices, including:

- Continual conformance with the International Organization for Standardization (ISO) standard for metadata registries -- ISO 11179-3:2003 and expansion of the registry in synchronization with the model;
- Expansion of the ISO 11179 metamodel to accommodate evolving technology advances;
- Incremental releases of the registry employing Carnegie Mellon's Capability Maturity Model for Software Level 3 processes to assess and improve software processes;
- The registry documents application metadata as well as standard and other well-formed data that conform to naming and definition guidance; and
- OEI's use of various approaches to solicit user input and to guide future development activities. These approaches include user conferences and a Web-based feedback mechanism for receiving comments.

1.8 Technical infrastructure

Application software is developed using Oracle Version 9i Data Base Management System (DBMS), Oracle Designer 9i, which generates Oracle Web modules, and dynamic HyperText Markup Language (HTML) procedures for the Web applications. This development methodology was selected to ensure a great deal of flexibility in development. The software is refreshed in new releases 3 to 4 times per year, and the Designer-generated software allows rapid response to changing requirements.

The software has been developed by systems engineering contractors certified in Carnegie Mellon's Software Engineering Institute (SEI) Capability Maturity Model

(CMM) Level 3. The application development method, combined with the SEI CMM Level 3 procedures and expert technical and subject matter staff, enables trouble-free deployment in EPA's technical infrastructure.